Terminology expansion and relation identification between genes and pathways

Dowdall, J ; Rinaldi, Fabio ; Persidis, A ; Kaljurand, K ; Schneider, G ; Hess, M

Abstract: This paper demonstrates the applicability of an NLP approach to medical articles that does not rely on the availability of an existing ontology. The analysis is synthetically exhaustive, progressing from flat, phrasal boundaries to hierarchical dependency relations between heads.

Terminology expansion and relation identification between genes and pathways

James Dowdall[*], Fabio Rinaldi[*], Andreas Persidis[†]
Kaarel Kaljurand[*], Gerold Schneider[*], Michael Hess[*]

Abstract

This paper demonstrates the applicability of an NLP approach to medical articles that does not rely on the availability of an existing ontology. The analysis is syntactically exhaustive, progressing from flat, phrasal boundaries to hierarchical dependency relations between heads.

1 Introduction

A gene contains hereditary information encoded in the form of DNA and is located at a specific position on a chromosome in a cell's nucleus. Genes determine many aspects of anatomy and physiology by controlling the production of proteins (gene products). Gene products form interconnected networks in order to accomplish specific goals. A biological process is accomplished by one or more ordered assemblies of molecular functions. Examples of broad biological process terms are "*cell growth and maintenance*" or "*signal transduction*". Examples of more specific terms are "*pyrimidine metabolism*" or "*alpha-glucoside transport*".

Understanding the relationships within and between these groups is central to biology research and drug design as they form an array of intricate and interconnected molecular interaction networks which is the basis of normal development and the sustenance of health.

One of the problems in this task is that current understanding of biology exists in islands of knowledge which are often ill connected. In recognition of this situation a number of approaches are currently being developed in order to help with the generation of hypotheses which can later be confirmed or refuted in wet lab experiments. Literature-based Discovery (LBD) is one such approach that uses free text (scientific articles) as its raw material.

The main problem being faced by LBD is that nomenclature being used by researchers is non-standard, polysemy, synonymy and homonymy are common resulting in significant problems. Ontologies attempt to provide a framework

[*]Institute of Computational Linguistics, University of Zürich, Switzerland. [†]Biovista, Athens, Greece.

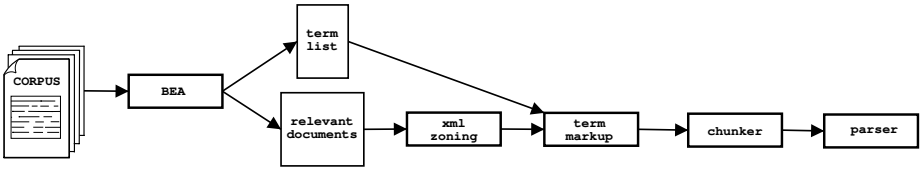

Figure 1: Processing stages

for common understanding of important concepts and their inter-relationships, but often they do not exist or if they do they are not widely accepted (see section 5). This creates a need for automatic creation of ontologies that will allow computer based systems to better understand and extract information from scientific articles.

2 The BioLab Experiment Assistant (BEA)

Two term lists, a gene list and a biological process (pathway) list, together with an initial corpus of scientific articles were collected using the BioLab Experiment Assistant tool from Biovista (see Figure 1).

To reduce the effect of polysemy and synonymy in the creation of the ontology but also to be able to check the accuracy of the creation process, a set of articles closely related to a specific set of genes and pathways was selected using the BEA tool. Given a set of research parameters (such as a set of genes and pathways) BEA returns highly relevant articles that can be used when evaluating the ontology extraction system.

The BEA identified 94 full research articles from MedLine. As these are available as html files the first requirement was to translate them into a more computationally friendly xml format. MedLine uses a uniform html template for all of the articles so the format transition was relatively straight forward. The freely available **html2text**[1] strips the html tags whilst preserving the basic document structure of sections, paragraphs and figures. Once non-ASCII characters are dealt with, simplistic processing translates this into the zones `<article>`, `<docinfo>`, `<title>`, `<sec>` and `<para>`. The `<docinfo>` contains the publication date, document id, the main title and the authors/affiliations.

This process is vital as it results in the ability to intelligently manipulate the document for further processing. Simply stripping the html tags to leave bare text yields a processable file but contains irrelevant zones (such as the bibliography). Also, some zone distinctions are lost altogether as in the case of captions on figures becoming part of their closest paragraph. XML zoning allows the syntactic analysis to be targeted more efficiently by not wasting the

[1] http://userpage.fu-berlin.de/~mbayer/tools/html2text.html

computational effort where there is nothing of interest to find.

The resulting xml documents contained 98512 words involving 37809 unique word forms.

3 Term Expansion

The two term lists identified by the BEA and used in article selection involved 4000 genes and 1300 pathways. The frequency of token length for each term list is shown in figure (2). The pathways display a canonical distribution of tokens. The most frequent being terms with two tokens, with the frequency steadily dropping as the term length increases. On the other hand, the genes are extraordinary in the concentration of single word terms.

The first step is to markup the terms identified by the BEA using additional xml tags (`<gene>` and `<pathway>`). This identified 900 genes and 218 pathways that occur in the corpus - represented as boxed tokens in figure (3). Next the entire corpus is chunked into nominal and verbal chunks using LT Chunk [4].

The chunker output represents the minimal phrasal groupings in each sentence. As such, it does not tackle the problems of prepositional phrase attachment or gerunds - this is left to the full parser (see section 4).

length	genes	pathways
1	3483	107
2	307	506
3	162	295
4	64	174
5	22	100
6	4	62
7	3	37
8	0	11
9	0	6
10	0	2
total	4000	1300

Figure 2: token frequencies

The corpus terms are then expanded to the boundary of the phrasal chunk they appear in. For example, NP3 in figure (3) contains two terms of interest producing the new term "*IFN-induced transcription*". The 1118 corpus terms were expanded into 6697 new candidate terms. 1060 involve a pathway in head position and 1154 a gene. The remaining 4483 candidate terms involve a novel head with at least one gene or pathway as a modifier.

Given the minimal nature of the chunks these expansions are relatively safe. FASTR [5] also exploits this safety margin but also crosses this boundary by arbitrarily attaching prepositional phrases and gerunds to known terms - resulting in an increase in noise. Our approach is to use a full syntactic parser to more accurately determine how these minimal chunks form a coherent sentence.

Using the sentence delimitation given by the chunker, 5718 sentences containing at least two terms were identified for further processing.

4 Parsing

The deep syntactic analysis builds upon the chunker using a broad-coverage probabilistic Dependency Parser [10], [11] to identify sentence level syntactic

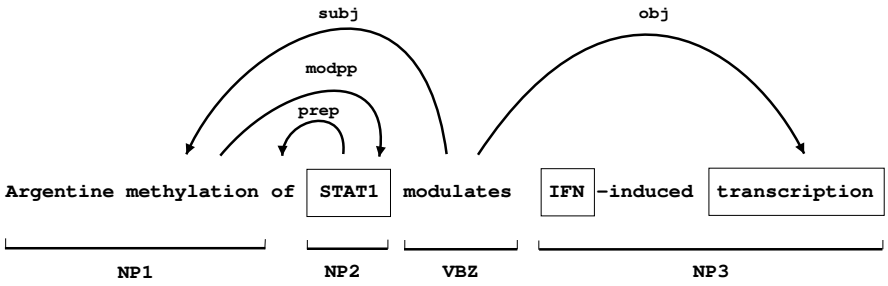

Figure 3: terms, chunks and dependency relations

relations between the heads of the chunks. The output is a hierarchical structure of syntactic relations - functional dependency structures - represented as the directed arrows in figure (3). The parser uses hand written declarative rules to encode acknowledged facts, such as verbs typically take one but never two subjects, combined with two probabilistic language models, similar to [2]. Both are supervised and are based on Maximum Likelihood Estimations (MLE). The first is based on lexical probabilities of the heads of phrases and calculates the probability of finding specific syntactic relations (such as subject, sentential object, etc.). The second probability model is a PCFG for the production of verb phrases. Although CFGs are alien to dependency grammar, verb phrase PCFG rules can model verb subcategorization frames which are an important component of a dependency grammar.

Figure (3) displays the three levels of analysis that are performed on a simple sentence. Term expansion yields `NP3` as a complete candidate term. However, `NP1` and `NP2` form two distinct, fully expanded noun phrase chunks. Their formation into a noun phrase with an embedded prepositional phrase is recovered from the parser's syntactic relations giving the maximally project noun phrase involving a term: "*Argentine methylation of STAT1*" (or juxtaposed "*STAT1 Argentine methylation*"). Finally, the highest level syntactic relations (`subj` and `obj`) identifies a transitive predicate relation between these two candidate terms.

The parser is robust in that it returns the most promising set of partial structures when it fails to find a complete parse for a sentence. So for sentences more syntactically complex than the illustrated example, as many hierarchical relations are returned as possible. This represents an advantage over dedicated shallow processing methods.

5 Related Work

Automatic knowledge extraction (or strategies for improving these methods) over Medline articles are numerous. These knowledge bases store linguistically or statistically inferred relations between objects.

For example, [3] identifies possible drug-interaction relations (predicates) between proteins and chemicals using a 'bag of words' approach applied to the sentence level. This produces inferences of the type: drug-interactions(Protein, Pharmacologic-Agent) where an agent has been reported to interact with a protein.

[12] uses frequently occurring predicates and identifies the subject and object arguments in the predication, in contrast [9] uses named entity recognition techniques to identify drugs and genes, then identifies the predicates which connect them. This type of 'object-relation-object' inference may also be implied [1]. This method uses 'if then' rules to extract semantic relationships between the medical entities depending on the MeSH headings they appear under. For example, if a citation has "*Electrocardiography*" with the subheading "*Methods*" and has "*Myocardial Infarction*" with the subheading "*Diagnosis*" then the former diagnoses the latter.

But where such linguistic inferences are stored in a KB as facts, statistical inferences are only used to visualize possible relations between objects for further investigation. [13] measures statistical gene name co-occurrence and graphically displays the results for an expert to investigate the dominant patterns.

The strategies for improving document retrieval (Pubmed) include methods for gene name recognition [8] and statistical tables of MeSH term co-occurrence [6].

There are three common themes among most systems that process medical articles. First, the body of the aricles are excluded, focusing attention on titles and abstracts. These are targeted due to their tendency to include terms and relatively uncomplicated syntax. So, secondly, any parsing is reserved to shallow NP identifying strategies [12] possibly supplemented with PP information [9]. Finally, the vast majority of research in this area is founded upon utilizing the UMLS MetaThesaurus[2]. Whilst this is fine for research purposes, the time lag between identifying novel genes and pathways and including them in the UMLS tools negates any competitive edge in discovering relations. By the time they are included in the UMLS they are 'old news' [7].

6 Future Directions

We are currently working on 1400 sentence parses in order to quantify performance. The initial indications are that the majority of errors can be corrected at the level of the chunker. A common chunking error is in splitting capitalized tokens into a separate chunk. This is unsurprising as it was developed for

[2] http://www.nlm.nih.gov/pubs/factsheets/umlsmeta.html

newswire where capitalized terminological symbols and acronyms are less frequent and of less interest. The solution is to process the corpus with an acronym finding program [14] and normalize capitalization.

The next phase will determine how best to exploit these syntactic dependency relations. Previous research has proved the utility of identifying top level relations, such as the subject and object of a verb (see section 5). Any deeper analysis quickly runs into the kinds of syntactic phenomena that requires additional parsing. The return for this effort is an increase in the number and type of syntactic relations that can be identified between terms. These can either be formally mapped onto an ontological structure or used as a fine grained measure of "syntactic distance" between terms in a sentence.

The logical extension of the approach is to investigate possible interaction with existing ontologies. There are two compatible directions: Ontology expansion through term and relation discovery or increasing the types of entities that are related to include the ontological terms. The obvious candidate for this is the UMLS MetaThesaurus but the emphasis will be on utilizing, rather than becoming dependent on, external ontologies.

References

[1] J. J. Cimino and G. O. Barnet. Automatic knowledge acquisition from medline. *Medical Informatics*, 1999.

[2] Michael Collins. *Head-Statistical Models for Natural Language Processing.* PhD thesis, University of Pennsylvania, Philadelphia, USA, 1999.

[3] M. Craven and J. Kumlien. Constructiong biological knowledge bases by extracting information from text sources. *Proceedings of the 8th International Conference on Intelligent Systems for Molecular Biology (ISMB-99)*, 1999.

[4] Steve Finch and Andrei Mikheev. A Workbench for Finding Structure in Texts. In *Proceedings of Applied Natural Language Processing*, Washington, DC, April 1997.

[5] C. Jacquemin. *Spotting and discovering terms through Natural Language Processing.* MIT Press, 2001.

[6] E. A. Mendonca and J. J. Cimino. Automated knowledge extraction from medline citations. *Medical Informatics*, 1999.

[7] Youngja Park, Roy J. Byrd, and Branimir K. Boguraev. Towards Ontologies on Demand. In *Proceedings of the Workshop on Semantic Web Technologies for Searching and Retrieving Scientific Data (ISWC-03)*, Florida, USA, October 20 2003.

[8] F. Proux, D.and Rechenmann, L. Julliard, V. Pillet, and B. Jacq. Detecting gene symbols and names from biological texts: A first step toward pertinent information extraction. *Genome Informatics*, 9, 1998.

[9] T. C. Rindflesch, L. Tanabe, J. N. Weinstein, and L. Hunter. Edgar: Extraction of drugs, genes and relations from the biomedical literature. *Pacific Symposium on Biocomputing*, 5:514–25, 2000.

[10] Gerold Schneider. A low-complexity, broad-coverage probabilistic Dependency Parser for English. In *Proceedings of NAACL/HLT Student session*, Edmonton, Canada, May 27-31 2003.

[11] Gerold Schneider. Extracting and Using Trace-Free Functional Dependencies from the Penn Treebank to Reduce Parsing Complexity. In *Proceedings of The Second Workshop on Treebanks and Linguistic Theories (TLT 2003)*, Växjö, Sweden, November 14-15 2003.

[12] T. Sekimizu, H. Park, and J Tsujii. Identifying the interaction between genes and gene products based on frequently seen verbs in medline abstracts. *Genome Informatics, Universal Academy Press.*, 1998.

[13] B.J. Stapley and G. Benoit. Biobibliometrics: Information retrieval and visualization from co-occurrences of gene names in medline asbtracts. *Symposium on Biocomputing, 529.540*, 2000.

[14] K. Taghva and J. Gilbreth. Recognizing acronyms and their definitions. *the International Journal on Document Analysis and Recognition, (IJDAR)*, 1:191–198, 1999.